GIS ГАЗИНФОРМ СЕРВИС

От референсной архитектуры полигона к архитектуре системы испытаний систем искусственного интеллекта

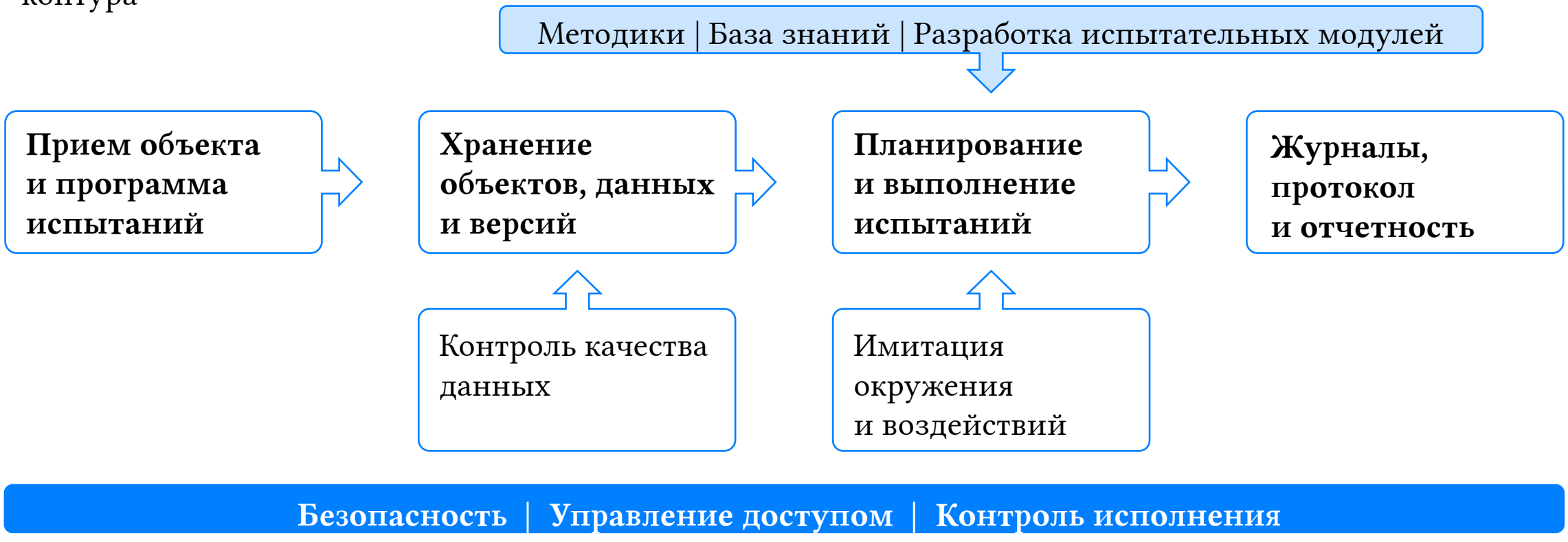
Как связать функции полигона с воспроизводимой процедурой и проверяемым основанием для заключения?

Виткова Лидия Андреевна

К.т.н., Начальник Аналитического центра кибербезопасности ООО «Газинформсервис», старший научный сотрудник ЛПКБ СПб ФИЦ РАН

Референсная архитектура задает функции полигона

Укрупненная схема исходного организационно-технического контура



Следующий шаг — определить связи между наблюдением, оценкой и итоговым заключением.

Объект испытаний — система в заданной конфигурации

Предлагаемая граница программно-методического ядра полигона

Заявитель
Объект и требования

→

Система испытаний
Процедура и свидетельства

→

Получатель
Решение о применении

$$C = \langle M_v, D_v, P_v, T_v, E_v \rangle$$

Mv — модель, версия

Dv — данные, версия

Pv — промпты и настройки, версия

Tv — инструменты и права, версия

Ev — среда и контекст, версия

Вывод не распространяется за пределы C и заявленного профиля применения с учетом указанных версий.

При изменении условий проверяют сопоставимость результатов.

Требование переводится в измеримую процедуру

GIS ГАЗИНФОРМ СЕРВИС

Пример — стойкость языковой системы к обходу заданных ограничений

Элемент процедуры	Что необходимо зафиксировать
Критерий и показатель	Ограничения, пороги, показатели (например из стандарта)
Задания и условия	<table><tr><th>Элемент процедуры</th><th>Что необходимо зафиксировать</th></tr><tr><td>Критерий и показатель</td><td>Ограничения, пороги, показатели (например из стандарта)</td></tr><tr><td>Задания и условия</td><td>Воздействия, запросы, конфигурация С</td></tr><tr><td>Оцениватель</td><td>Правило разметки ответов и версия оценивателя</td></tr><tr><td>Наблюдение и метрика</td><td>Успех воздействия; доля по валидным испытаниям</td></tr></table>
Оцениватель	Правило разметки ответов и версия оценивателя
Наблюдение и метрика	Успех воздействия; доля по валидным испытаниям

Предлагаемая архитектура системы испытаний

GIS ГАЗИНФОРМ СЕРВИС

Функциональные компоненты связаны процедурой получения свидетельств

Постановка испытания
- → Паспорт объекта
- → Требования и методики

Исполнение заданий
- → Оркестрация
- → Воздействия и сценарии

Оценивание и решение
- → Измерительная цепь
- → Калибровка оценивателя
- → Неопределенность и правило

Пакет свидетельств
- → Наблюдения
- → Протокол и вердикт

Кампании | Контрольные точки | Происхождение и версии артефактов

Сквозная связь: объект → методика → прогон → оценка → правило решения → свидетельство

Прототип MLSecOps

Демонстрация MLSecOps SOMA

GIS ГАЗИНФОРМ СЕРВИС

Наблюдаемая функция

- → Карточка модели
- → Ресурсы и связи
- → История версий данных

Вопрос апробации

Можно ли восстановить полную конфигурацию конкретного прогона?

Наличие карточки — наблюдаемая функция; полнота паспорта и воспроизведение условий требуют проверки

Прототип режима анализа защищенности в MLSecOps

GIS ГАЗИНФОРМ СЕРВИС

Наблюдаемая функция

- → Подключение цели
- → Профиль испытания
- → Матрица покрытия

Матрица описывает показанный прогон; общий вывод требует заданий, знаменателя и проверки оценивателя

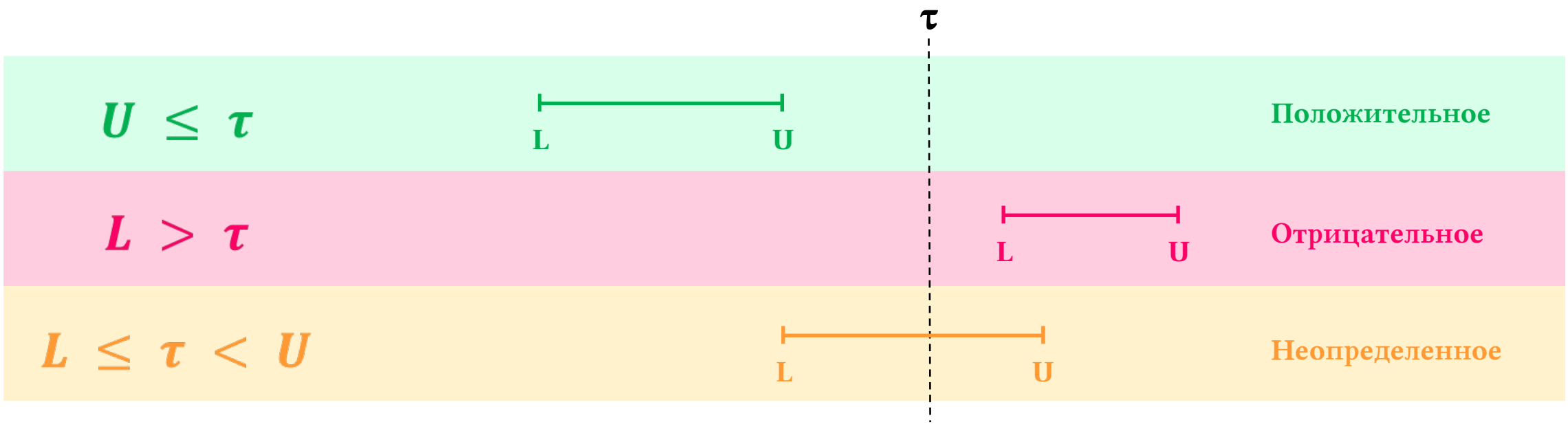
Заключение требует оценки неопределенности

GIS ГАЗИНФОРМ СЕРВИС

Оценка в интервале

$$I = [L, U],$$

τ — порог, заданный программой испытаний



Источники неопределенности: повторы и выборка; ошибки оценивателя; неоднозначность определения.

Схематический пример без численных данных; правило выбирают до испытаний и обосновывают для применения.

Мониторинг во время эксплуатации

GIS ГАЗИНФОРМ СЕРВИС

Наблюдаемая функция

- → Качество и дрейф
- → Различия моделей
- → Атрибуция признаков

Вопрос выхода на повторное испытание?

Сигнал → кампания → сопоставимое повторное испытание

Планируемый результат — архитектура проверяемого заключения для СИИ

От функций полигона к связям между объектом, процедурой и свидетельствами.

Методический результат	Выделены функциональные носители оценивания, решения и свидетельств
Практический результат	Показаны отдельные функции учета, исполнения и наблюдения изменений
Следующий научный шаг	Сквозная апробация воспроизводимости и надежности оценивания

Вывод ограничен конфигурацией объекта, методикой и условиями испытаний

ГАЗИНФОРМ СЕРВИС

Вопросы

Виткова Лидия Андреевна

К.т.н., Начальник Аналитического центра кибербезопасности ООО «Газинформсервис», старший научный сотрудник ЛПКБ СПб ФИЦ РАН